Market Regime Forecasting using Correlation Networks and Machine Learning Classifiers

Fan Zhu

Honor Thesis

Economics Department

University of North Carolina at Chapel Hill

Thesis Advisor: Dr. Michael Aguilar

Faculty Advisor: Dr. Jane Fruehwirth

May 1, 2020

Approved by:

Dr. Michael Aguilar

Dr. Jane Fruehwirth

Market Regime Forecasting using Correlation Networks and Machine Learning Classifiers

Dr. Michael Aguilar* Fan Zhu†

May 1, 2020

Abstract

Market regime classification has been practically influential to financial practitioners. Through the top-down approach to identify market regimes, we partition historical data on U.S. equity market returns into bear and bull markets (Hanna, 2018). Then we create a correlation network for each market state respectively. The topological features of those networks will be fed into the machine learning classifiers, which will provide a model whose machine learning algorithm generates the highest accuracy among all the candidate algorithms. We found that, with 1-months duration and 1-week lagged industry return data, the Fine Tree model has the highest forecasting accuracy of 91.7%. For data with longer horizon, the Kernel Naive Bayes model with 3-month duration and 12-month lagged data performs the best with an accuracy of 66.7%. The applications of the acquired model can not only benefit financial practitioners from acting in advance of the market, but also benefit macro-economists from preparing better for the extreme market events.

Keywords: Market Regime, Correlation Network, Machine Learning

*Correspondence author/Thesis advisor: maguilar@email.unc.edu
†First author: zfan20@live.unc.edu

1 Introduction

Detecting and forecasting regime shifts has been a crucial topic for decision makers in a dynamically changing environment. Every financial professional prefers to ride on the rising tide of a bull stock market and minimize their loss during a bear market. Monetary policy makers aim to shrink the size of an inflating bubble and implement the defibrillator to a plunging market. To act in advance of the market turning points, practitioners and researchers have put considerable merits and efforts in the detection and forecasts of market regimes (Diebold and Weinbach, 1993; Lunde and Timmermann, 2004; Kole and Dijk, 2010).

The complexity of the stock market complicates market forecasting. One of the most significant features of this complex system is the interactions between stocks, which are commonly captured by the correlation coefficients (Mantegna, 1999). The theoretical pavement for using the correlation to fathom the market structure is that the price of a financial asset is influenced by both macroeconomic factors and the prices of other assets (Fama and French, 1992). The change of the economic factors and the assets' prices would result in an information flow that eventually affect other assets. Such interconnection between financial assets constitutes the fundamental structure of the market. To obtain an efficient representation of the market structure, many researchers have turned to the network analysis, which has been extensively used in the fields of biochemistry and neuroscience (Kazemilari and Djauhari, 2015). The network theory for the financial market basically states that each node in the network represents a single asset, and the edge that connects each pair of assets suggests the correlation between the prices of these two assets (Guo et al., 2018). Degiannakisa and Florosc (2013) found that the correlations between European industrial sector portfolio returns and the oil price returns vary over time and are specific to each industry. Based on this finding, I hypothesized that the correlations between industry indices returns may also change over time, which may reflect on the general structure of the stock market. Therefore, my thesis will contribute to previous work by utilizing the time-varying correlations between industry returns to predict market regimes.

The applications of the model and methods in my thesis have the potential of generating enormous profits in the field of asset management. For instance, simply incorporating different market conditions into the asset allocations will yield higher portfolio returns and lower portfolio risks than the conventional mean-variance portfolio optimization methods (Bernhart and Zagst, 2011). The implementation of machine learning classification algorithms tends to generate higher forecasting power, which may lead to even better portfolio performance. We can also replace the industry portfolios with international stock portfolios to capture the dynamics of the international stock markets. Ang (2002) found that combining international diversification with regime-shifting models could still be profitable even in the bear market regimes with higher volatility. With correlation networks for the markets of interests, researchers could then apply the machine learning model in this paper to predict when the trough of the market will arrive and which side of the trade they will choose. Based on these forecasts, more efficient trading strategies and economics policies could be implemented in advance to counteract or even benefit from the upcoming market plunge.

In this study, we will investigate whether the topological features in the correlation networks constructed from market data can be utilized for forecasting market regimes. We will begin with creating correlation networks for the overall 49 industries of the U.S. economy (French, 2019). Then the quantitative attributes of these networks will be extracted to predict corresponding market regimes. Our paper ends with a short-term Fine Tree model with an accuracy of 91.7% and with a long-term Kernel Naive Bayes model with an accuracy of 66.7%.

This paper proceeds as follows: Section II will provide a thorough literature review on the topics and techniques that will be covered in this paper. Section III will provide empirical models for the market regimes classification algorithms, minimum spanning tree algorithms and candidate machine learning classification algorithms. Section IV will scrutinize the data used in this research. Section V will present and interpret the descriptive statistics and the results from the machine learning classification models. Section VI will discuss the summary

of this research, potential improvements and major contributions to the existing literature.

2 Literature Review

Correlation networks have been established as an efficient method to investigate the market structure. Mantegna (1999) is among the trailblazers who applied the network theory to transform the correlations between stocks into a hierarchical taxonomy. Mantegna converted the Pearson's correlation coefficient matrix of stocks in the Dow Jones Industrial Average index (DJIA) portfolio into the Euclidean distances between any two stocks, which were then fed into the minimum-spanning-tree algorithm to generate a visualization of the hierarchical arrangement of the market. The results of the hierarchical trees of stocks indicated the viability of using the network theory to model the underlying systems for the financial market.

Kullmann et al. (2002) has shown the viability of utilizing lagged correlations between assets to forecast the asset returns. They investigated the pulling effects between stocks by analyzing the lagged cross-correlation between the returns of stocks at the New York Stock Exchange. Kullmann and his colleagues based their methodology on two types of mechanisms that elucidate how stocks are correlated: 1. External effects such as political and economic events, which affect both stock prices simultaneously. Thus, the maximum of the correlation between these two stocks is at zero-time shift. 2. One of the companies has an effect to the other's stock price. The change of the influencing stock's price would result in a change of the influenced stock's price in a later time since the influenced stock requires some time to react in its price. Their results showed that the characteristic time shift, identified by the position of the maximum correlation, was usually a few minutes after the order had been placed, which is consistent with the efficient market hypothesis. Kullmann et al. (2002) also found that, with more trades, the stock prices of more important companies pull those of relatively smaller companies. These studies confirmed the validity of utilizing correlation

networks to study the interactions between stocks. Analogous to Kullman's research, my study will apply the network analysis to capture the interactions in the stock market and how such inter-activities forecast the market states.

Similar to studying correlations between each asset, the interactions between different markets can also be investigated through the correlation networks. For instance, Sun et al. (2019) studied the interactive systems between the sovereign credit default swaps (CDS), stock and commodity markets through a spillover correlation network. Their dataset includes sovereign CDS spreads, stock indices, and commodities from developed countries (G7 countries) and developing countries (BRICS countries), with a time period from 2009 to 2017. To estimate the spillover effects for stock returns, the researchers resorted to the forecast error variance (FEV) decomposition of the rolling average of the Vector Autoregression model, whose results were then utilized to generate the correlation networks for the spillover effects between stocks. Sun and his colleagues discovered that the effects from the sovereign CDS to the stocks in developing countries exhibited a greater magnitude than those in the developed countries. This finding broadens the possibilities of the basic elements in the financial network analysis from stocks to more aggregate entities, such as an asset market or, in this paper, an industry.

In order to input the correlation networks for the forecasting algorithms, we need to extract the representative quantitative features of the correlation networks. Bonanno et al. (2003) studied whether the CAPM or the random model can describe the topological properties of the MST constructed from the stock market data. To compare the minimum spanning trees, they extracted the distributions of degrees and the in-component degrees from each network. The comparisons showed that both the random and the one-factor model failed to capture the real market's hierarchical distribution of the stocks. The MST from the random model shows a nonhierarchical structure, while the MST from the one-factor model exhibits a hierarchy with only one center. These results emphasized the complexity of the real market's structure to which the simple models such as the random model and the one-factor model

failed to approximate. Therefore, to properly characterize the complexity of the market, we will extract similar topological properties used in Bonanno's study.

As the last step in this thesis, forecasting the market regimes has been widely researched. Upadhyay (2012) predicted the categories of the Indian financial markets through incorporating different financial ratios into the multinomial logistic regression model. Their prediction results exhibited an accuracy of 56.8%. As the machine learning has become popular in financial research, Pierdzioch and Risse (2018) implemented one of the machine learning algorithms, the boosted regression trees, to choose the set of predictors that can maximize the forecasting power of the model. Pierdzioch's study presents evidences that support the rational expectations hypothesis (REH) for short-term stock market predictions and evidence that reject REH for longer-term market forecasts. In my thesis, I will fill the blank of linking network analysis with machine learning forecasting models. By comparing the accuracy of our model with that of other forecasting models, I will test whether our machine learning model has an comparative edge among other market forecasting models.

Based on the above-mentioned prior literature, this thesis will contribute to empirical finance by feeding industry-level data into the network analysis and implementing machine learning algorithm to forecast market regimes. Because the industry-level data is the aggregate stock returns within the specific industry, we expect that their time-series behaviors and the structure of the resulted correlation networks tend to exhibit seasonal or cyclical patterns that last from one quarter to longer than one year. Following this reasoning, we expect that forecasting models with input data of proper duration and lags would generate the highest accuracy.

3 Empirical Model

3.1 Market Regime Identification

To classify the market regimes, we adopted the top-down approach proposed by Hanna (2018). Hanna established five principles for any methods of market regime classification:

1. There exists an alternating pattern of the bear and bull regimes.
2. There is a significant positive/negative total return for each bull/bear market state.
3. The prices during each market regime should be bounded by end point value of the corresponding regime period.
4. A minor change in the parameterization would not result in significant changes of dates of each regime period.
5. Extending the time period of the time-series data would not result in significant changes of the dates of each regime period.

This top-down method is based on the identification of local extrema, which serve as the candidate turning points between different market states. Applying the above-mentioned five principles to this basis would result in three concrete steps as shown below:

- Pre-phase: Locate the left maximum and right minimum as the reversal points in the price series $P(t)$ in a time interval $[t_a, t_b]$. Hanna defined the reversal points as below
  - left maximum as points t_i such that $P(t_j) \leq P(t_i)\ \forall j < i$
  - right minimum as points t_i such that $P(t_i) \leq P(t_j)\ \forall i < j$
- Phase 1: Partition the intervals between each reversal points into sub-intervals recursively until no new partition can be made based on the partition algorithm.

- locate the absolute minimum and maximum values within each interval $[t_a, t_d]$. Without loss of generality, we denote P_{t_b} as the minimum and P_{t_c} as the maximum, which assume that $t_b < t_c$.
- When there exist repetitive extreme values, we select the value which occurred at a chronologically later time.
- If $[t_b, t_c] \subsetneqq [t_a, t_d]$, and either $P_{t_c}/P_{t_b} < 1 - \lambda_{bear}$ or $P_{t_c}/P_{t_b} > 1 + \lambda_{bull}$, then partition the interval $[t_a, t_d]$ at t_b and t_c.

- Phase 2: Partition each sub-intervals recursively based on the following rule until no new partition can be made
  - If the maximum valid reversal over $[t_b, t_c]$ is found within each (sub)-interval $[t_a, t_d]$, then partition $[t_a, t_d]$ into three sub-intervals $[t_a, t_b]$, $[t_b, t_c]$ and $[t_c, t_d]$.

The general rules are described above following the Hanna (2018)'s method. One complementary note about the market extremes is that any market trend that lasts less than 20 days is classified as either a rising market rally or a declining market correction.

3.2 Correlation Network

We chose to use the minimum spanning tree (MST) for our correlation networks. MSTs are a well-established method of forming networks from similar data. For our networks, we used Matlab's graph and minimum spanning tree functions to construct them. The specific methodology used to create these trees is as follows (Mantegna and Stanley, 1999):

1. First, we obtain the daily adjusted closing price for each stock and calculate the log returns.

2. Then we compute the correlation coefficient C_{ij} between each pair of stocks and obtain a n-by-n matrix of C_{ij}. The subscripts i, j and k represent different financial assets.

3. We convert the correlation coefficients to Euclidean distances representing the edges in the network. Each edge distance must satisfy three axioms:

   - $d_{ij} = 0$ if and only if $i = j$
   - $d_{ij} = d_{hi}$
   - $d_{ij} \leq d_{ik} + d_{kj}$

4. However, the direct application of the correlation coefficient does not satisfy these axioms. Thus, we need to transform these correlation coefficients to be qualified for the three axioms. One of the possible transformation functions is

$$d_{ij} = \sqrt{2 \cdot (1 - C_{ij})}$$

5. There are several feasible algorithms to determine the rules of forming links from the Euclidean distance to a correlation network. One of them is Kruskal's Stress-1 score, which is a well-established measure for the goodness of fit in a graph model (Kruskal, 1964). The Kruskal's stress score is computed as the formula shown below. The d_{ij} represents the distance between the industry i and industry j. The δ_{ij} means the disparity between industries i and j. The distances and disparities are calculated based on the rules described previously. As a result, the graph with the lowest Kruskal's Stress-1 score is the most efficient representation of the minimum spanning tree. Higher number of points in the graph would lead to larger stress score.

$$Stress = \sqrt{\frac{\sum (d_{ij} - \delta_{ij})^2}{\sum d_{ij}^2}}$$

The resulting correlation network can be classified as an undirected minimum spanning tree, because the correlation coefficients are transformed into Euclidean distances that are all positive and directions of influences cannot be traced. To characterize this undirected

network properly, we will extract the following useful topological properties: stress score, edge weights and centrality. Each node in the minimum spanning tree is equivalent to one industry. The edge, which is the line connecting each pair of nodes, represents the correlation between the paired industry returns. The longer edges indicate smaller magnitudes of correlations and vice versa. The centrality of each node provides information about the relative location of each node in the system. The absolute coordinates do not matter because the the undirected network is a three-dimensional representation, which can be spatially rotated.

Beside the topological characterizations of the networks, we also include the properties of the S&P500 returns in the set of predictors. Wee and Yang (2011) found that during a bull market, the market volatility would increase while the liquidity will increase. Thus, we incorporate the standard deviation, kurtosis and range of the market returns into the forecasting model.

To forecast market regimes, we will use 'lagged' data in our training sample. For instance, '1-month lag' indicates that we will use correlation network data to forecast the market type for one month in advance. We will attempt the lags of one month, three months, six months and twelve months, because we hypothesized that the model with seasonal lags would be a better fit for the general business cycle for industries, which would yield a higher accuracy rate. For relatively short-term forecasting, we will also attempt the lag of one week, two weeks, three weeks and four weeks.

The 'duration' of the training data represents the size of the estimation window. For example, a data duration of 'one month' indicates that to forecast each market turning point, we will use an estimation window of one-month industry-level data. The smaller estimation window size may capture the sudden change of the market structure better, while a large estimation period size will provide more information to construct a more accurate long-term correlation network. Thus, we will attempt the data duration of one month, three months, six months and twelve months. As a result, we will attempt 16 combinations between different data lags and duration.

3.3 Machine Learning Models for Forecasting Market Regimes

To build the model for predicting market regimes, we utilize the Matlab's Classification Learner App, which applies a number of different machine learning classification models to the samples drawn from the correlation networks. To prevent the models from overfitting, we will apply the 10-fold cross-validation method by partitioning the dataset into 10 folds, training the model based on 9 folds and testing the accuracy of the model on the remaining one fold. The input data for the machine learning models include quantitative characterizations of the correlation networks for each market state. Specifically, the features comprise the degree distributions, in-degree components, stress values and the densities from each network (Bonanno et al., 2003). Then we will employ the classification app to determine if there are any patterns before, during, or after market regime switches, recurring often enough to be used to classify, or even forecast when switches will occur, and what it will switch to.

The most common models used are briefly described below. This is followed by a short explanation on how the models are kept from being overfitted, or being too closely linked to the training data, thus becoming unable to generalize for new data.

3.3.1 Nearest Neighbors (KNN)

Nearest Neighbors, or K-Nearest Neighbors, is a non-parametric lazy learning algorithm. It makes no assumptions about the distribution of the data, instead determining that information from the data itself, hence its non-parametric characteristic. KNN also has essentially no training period, instead taking new data and comparing it to the "memorized" test data, and classifying it then, making it a lazy algorithm because it does not formulate a classifying model equation during a traditional testing period. It simply holds all of the input data, and then adds to its collection.

Data in KNN are classified usually by a majority vote of its k-nearest neighbors, joining whatever class it most closely resembles. KNN is a simple to understand, versatile classifier

with high accuracy, but suffers from high memory costs, slow predictions, and over-sensitivity to noise and size.

3.3.2 Decision Trees

Decision trees are a classification and regression tool that choose features in the provided data to "split" data between, branching downwards like an upside-down tree. They are some of the simplest machine learning outputs to understand and visualize, require little-to-no data prep from the user, and implicitly use feature selection.

3.3.3 Ensemble Classification

Ensemble Classification is a machine learning classification technique that utilizes a number of different models to create a single, hopefully superior, model. There are a number of different ensemble methods, including bagging (bootstrap aggregating), random forest, boosting, and stacking. Bagging simply creates multiple models using different data samples drawn with replacement (bootstrap sampling), and then averages their results. Random forest is similar to bagging, using bootstrap sampling to create each tree's dataset, except each different tree is given only random subset of all the available features that can split upon, rather than having all features available. Boosting takes weak classifiers, and retrains them on weighted versions of the training data, until they become accurate. Stacking combines multiple classifiers through a meta-classifier. The base models are trained on the complete training data, and successive models are trained on the previous model's outputs.

3.3.4 Support Vector Machines (SVM)

Support Vector Machines is a discriminative classifier that, given labeled training data, outputs a hyperplane that separates new data into the two classes defined by a subset of the data, called support vectors. If the data is not linearly separable, SVM projects the system to higher and higher dimensions until it can be linearly separated.

SVM models are built by separating data into two subsets, the larger used to identify support vectors, and the smaller one to test accuracy. The main advantage of SVM is its ability to properly identify global, rather than local, minima, which allows SVM to generalize easily to new data

3.3.5 Overfitting

In a machine learning scheme, a common problem for models is overfitting, when the model is too sensitive to noise or short-term only effects. This means that while the model may be accurate for the training data, it is too accurate, and cannot easily generalize to new data. To avoid this problem, we utilized Matlab's 10-fold cross validation. A classic k-fold cross validation splits the dataset into k subsets, and repeats the training-testing steps k times. Each time, a different subset of the data is used as the testing data, and the remaining k-1 subsets are used as the training data.

4 Data

The S&P500 daily returns are pulled from FactSet as the market index, which is fed into the heuristic for classifying market regimes based not only on magnitude of returns, but also on overall duration of majority positive or negative returns. The daily frequency is required to achieve the precision of partitioning the historical market returns into different market types (Hanna, 2018).

The daily returns of the 49 U.S. industry portfolios are obtained from the Kenneth French's Data Library (French, 2019). The frequency of the industry portfolio data is consistent with the frequency of the market index data so that we can use the turning points of the market to partition the industry-level data. The earliest date with the complete cross-sectional data of 49 industry daily returns is 1978/03/01. Therefore, to maintain a balanced panel data, I synchronized S&P500 daily returns with the 49 industry daily return data so

that the entire estimation period is from 1978/03/01 to 2019/06/28.

5 Results

5.1 Market Regime

Through our market regime classification algorithm, we classify the daily S&P500 (1978/03/01-2019/06/28) into 30 bull markets, 26 bear markets, 5 market upward rallies and 7 market downward corrections. Table 1 shows the frequency of four market regimes and Figure 2 shows the partitioned S&P500 historical prices. The green areas indicate the bear markets, while the red regions show bull markets. The table in the appendix demonstrates the specific partition dates for the peaks, troughs, rallies and corrections.

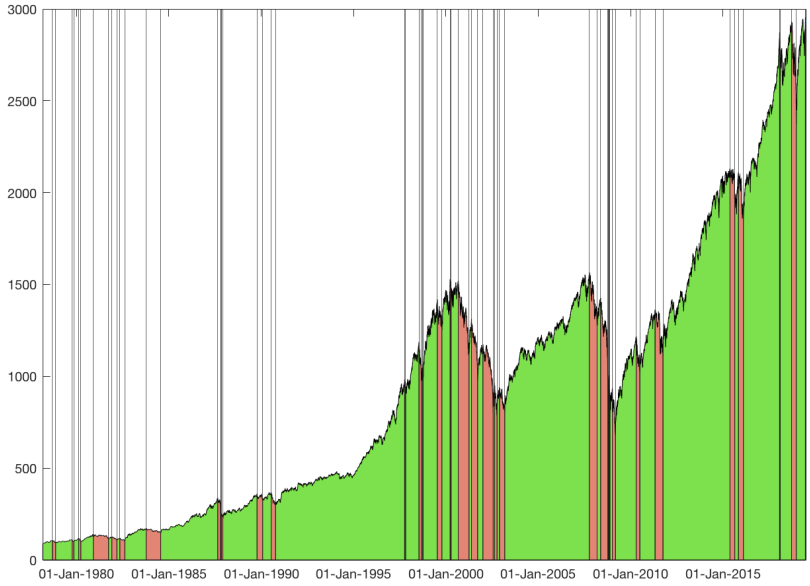
Table 1: Number of Market Regimes

Regime	Bear	Bull	Rally	Correction
Number	30	26	5	7
Frequency	44.12%	38.24%	7.35%	10.29%

Table 2: Descriptive Statistics for Market Regime Duration (Unit: Day)

Regime	Peak	Trough	Rally	Correction
Mean	286.3103	77.1923	6	10.5714
Standard Deviation	398.7468	49.1626	5.9582	3.9521
Skewness	2.3536	1.3335	1.0704	-0.8981
Kurtosis	8.2381	4.1830	2.6984	3.0444
Minimum	22	23	1	3
25^{th} Percentile	49	42	1.75	9.25
50^{th} Percentile	132	66.5	5	11
75^{th} Percentile	286.5	101	8.5	13.5
Maximum	1767	208	16	15
Range	1745	185	15	12
Number of Regime	29	26	5	7

Figure 1: S&P500 Historical Prices



The Table 2 showed the descriptive statistics for the market regime duration. According to the table, the bull markets (M = 286.31 days, SD = 398.75 days) have the longest average duration, while the market rallies (M = 6 days, SD = 5.96 days) have the lowest average duration. The bull markets exhibit the highest standard deviation in its duration, while the duration of the market corrections have the lowest standard deviation. The minimum duration for the market rally is only 1 day, which reflects extreme market movement within one day.

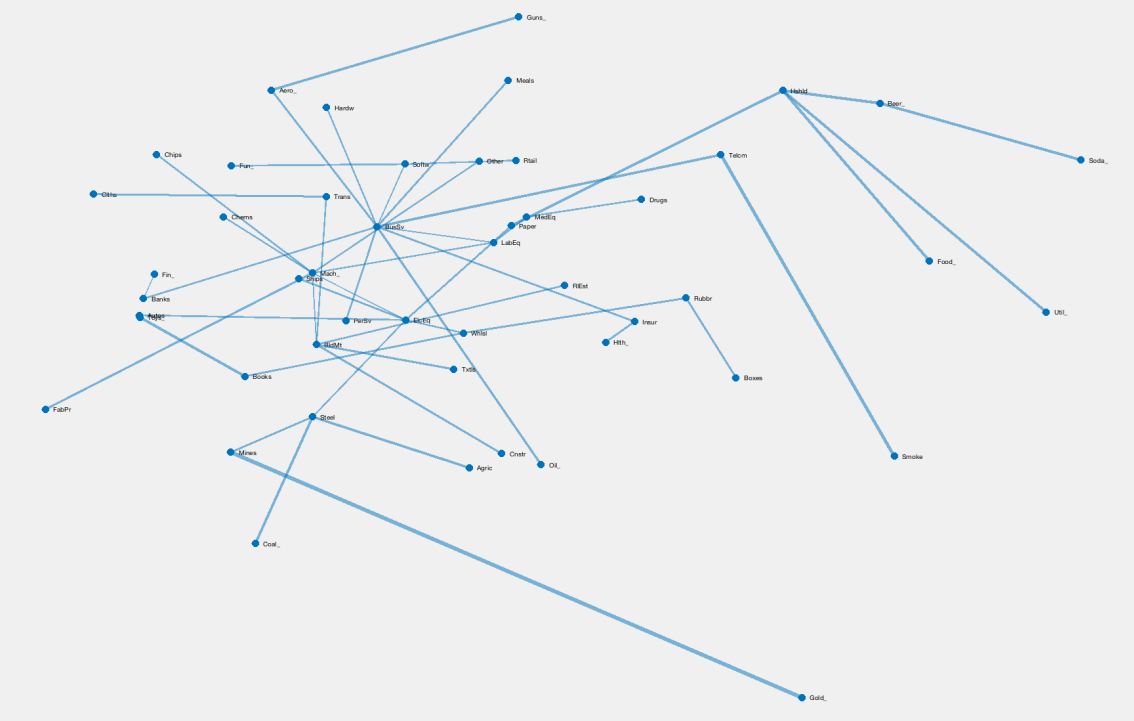
The Table 3 showed that the bull markets (M = 44.12%, SD = 47.18%) have the highest average total return, while the bear markets (M = -19.06%, SD = 7.32%) have the lowest average total return. The bull markets exhibit the highest level of risks while the market rallies have the lowest risks. The total return distribution of the bull markets is the most skewed to the left and the bull markets' distribution showed the largest kurtosis, which indicated that it had the fattest tails among the three distributions.

Table 3: Descriptive Statistics for S&P500 Total Return within Each Regime (Unit: %)

Regime	Peak	Trough	Rally	Correction
Mean	44.1202	-19.0573	13.7352	-13.5162
Standard Deviation	47.1776	7.3167	3.0035	5.4586
Skewness	2.5211	-0.9703	0.8463	-1.6355
Kurtosis	10.0038	2.9984	2.1627	4.1458
Minimum	11.3005	-36.9689	11.3666	-25.1862
25^{th} Percentile	17.5039	-19.9187	11.5267	-14.5158
50^{th} Percentile	22.4505	-17.8571	12.3336	-11.1885
75^{th} Percentile	57.9704	-13.5446	15.8098	-10.3185
Maximum	232.7422	-10.2335	18.4741	-10.0038
Range	221.4416	26.7355	7.1075	15.1823
Number of Regime	29	26	5	7

5.2 Correlation Network

Figure 2: Correction Network from sample S&P500 data (2018/11/28-2019/05/22)



The correction network in figure 2 visualizes the correlations between the returns of 49 industries in United States. Each node represents one industry. The longer the edge, the

lower the correlation between these two industries' returns. From the graphs, we can observe that some nodes have more connections to other nodes, which implies higher centrality of those industries. An industry with higher centrality in the correlation network has a larger far-reaching effect on other industries. For instance, the industry with the highest centrality in this correlation network is the business services industry, which had the highest number of significant correlations with other industry. In contrast, industries on the border of the network, such as the gold and gun industries, had less impact on other industries during the sample period.

Because the numbers of market rallies and corrections are too small for a valid sample size, we exclude market rallies and corrections from the input data for Machine Learning Classification Algorithms. Therefore, for each bull and bear market, we construct their corresponding correlation networks. As a result, there are in total 53 correlation networks in the input data for the machine learning classification algorithms.

5.3 Machine Learning Classification Results

The four summary tables below demonstrated the forecasting results of machine learning models using different duration and lags in the input data. In short-term forecasting, using 1-month duration and 1-week lagged data will provide the highest forecasting accuracy of 91.7% with the Fine Tree model. In the long-term forecasting, utilizing 3-month duration and 12-month lagged data generates the highest accuracy of 66.7% with the Naive Bayes model.

For the models with input data of 1-month duration, the greater lag results in lower forecast accuracy. The greater duration of the input data does not lead to higher forecast accuracy. The models with input data of 1-month lag tend to have greater accuracy except for the model with 3-month duration and 12-month lag input data.

The confusion matrix of the Fine Tree model shows that the model is slightly better identifying bull markets than identifying bear markets (Figure 3). In the graph, '1' indicates

Table 4: Machine Learning Classification Results for 1-Month Duration Data

Input Data	1-month duration 1-week lag	1-month duration 2-week lag	1-month duration 3-week lag	1-month duration 4-week lag
Accuracy	91.7%	77.1%	62.5%	62.5%
Model	Fine Tree	Cosine KNN	Subspace Discriminant	Naive Bayes

Table 5: Machine Learning Classification Results for 3-Month Duration Data

Input Data	3-month duration 1-month lag	3-month duration 3-month lag	3-month duration 6-month lag	3-month duration 12-month lag
Accuracy	62.5%	52.1%	58.3%	66.7%
Model	Fine Tree	Boosted Tree	Linear Discriminant	Naive Bayes

Table 6: Machine Learning Classification Results for 6-Month Duration Data

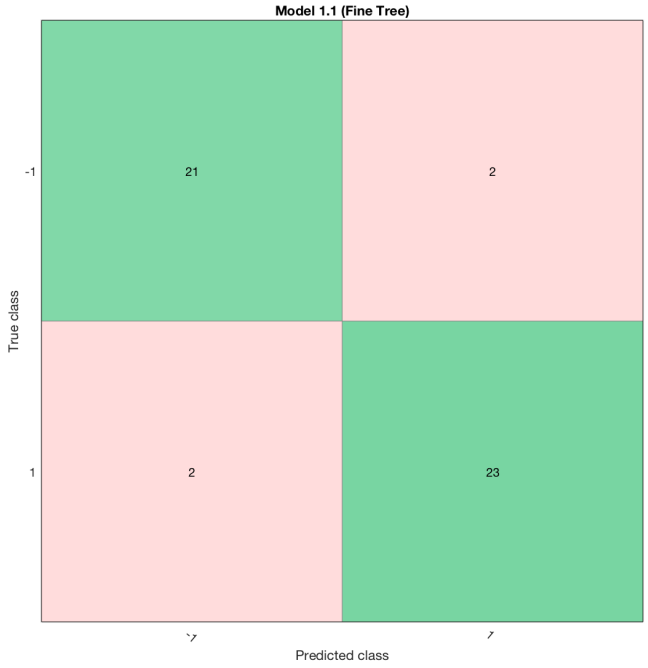
Input Data	6-month duration 1-month lag	6-month duration 3-month lag	6-month duration 6-month lag	6-month duration 12-month lag
Accuracy	64.2%	54.7%	58.3%	52.1%
Model	Linear Discriminant	Cubic SVM	Fine Tree	Coarse KNN

Table 7: Machine Learning Classification Results for 12-Month Duration Data

Input Data	12-month duration 1-month lag	12-month duration 3-month lag	12-month duration 6-month lag	12-month duration 12-month lag
Accuracy	54.7%	52.8%	54.2%	52.1%
Model	Coarse Gaussian SVM	Coarse KNN	RUSBoosted Tree	Coarse KNN

the bull market while the '-1' represents the bear market. The green color indicates the successful forecasts, while the red means failure. In a total sample of 23 bear markets, the model successfully identified 21 bear markets and failed to identify 2 bear markets. In a total sample of 25 bear markets, the model successfully identified 23 bull markets and failed to identify 2 bear markets. The model accuracy for forecasting the bear market is higher than the accuracy for the bull market. The accuracy of this model is 91.7%, which is the highest among all other models with various combinations of data duration and lags.

Figure 3: Confusion Matrix for the Fine Tree Model



The Figure 4 demonstrates the confusion matrix of the Kernel Naive Bayes model which is better identifying bear markets than identifying bull markets. In a total sample of 23 bear markets, the model successfully identified 17 bear markets and failed to identify 6 bear markets. In a total sample of 25 bull markets, the model successfully identified 15 bull markets and failed to identify 10 bear markets. The model accuracy for forecasting the bear market is higher than the accuracy for the bull market. The accuracy of this model is 66.7%, which is the highest among all other models with various combinations of data duration and

lags.

Figure 4: Confusion Matrix for the Kernel Naive Bayes Model

The Receiver Operating Characteristic (ROC) curve shows how much the linear discriminant model is capable of distinguishing between bull and bear market regimes (Figure 4). The y-axis of the ROC curve is the true positive rate and the x-axis is the false positive rate. The larger the Area under Curve (AUC), the more accurate the model's forecasts. In other words, the more the ROC curve deviates from the diagonal line, the high model accuracy. If the ROC curve is closer to the diagonal line, then it is more likely that the model forecasts are due to pure chance. Thus, Figure 5 and 6 show that the Fine Tree model with 1-month duration and 1-week lagged data performs much better than the Kernel Naive Bayes model with 3-month duration and 12-lagged data.

Figure 5: Receiver Operating Characteristic (ROC) curve for the Fine Tree Model

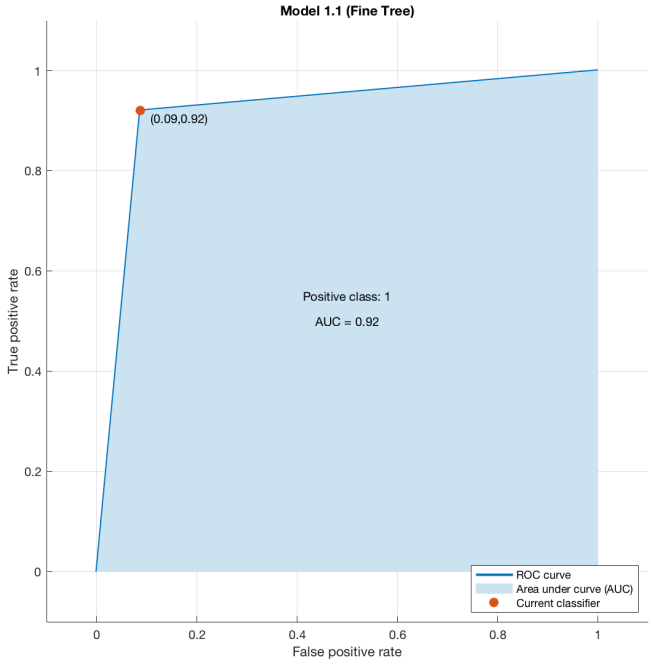
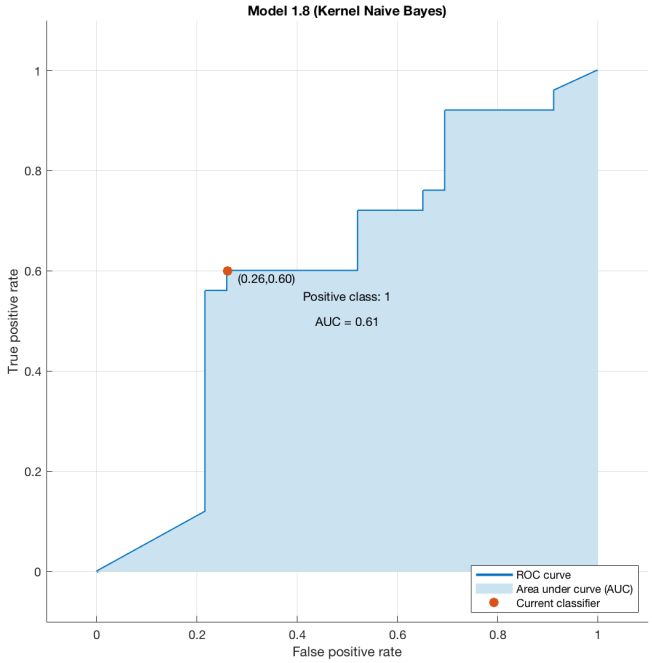


Figure 6: Receiver Operating Characteristic (ROC) curve for the Kernel Naive Bayes Model



6 Discussion

The results shows that, in the short-term, using the 1-month duration and 1-week lagged industry return data, the Fine Tree model has provided a model with the highest accuracy of 91.7% among other candidate models. In the long-term, model accuracy tend to decrease. That is, as the lag of input data increases, the model accuracy generally decreases. The most accurate (66.7%) long-term model is the Kernel Naive Bayes model with input data of 3-month duration and 12-month lags. The results of the various models in the results section do not completely confirm our hypothesis that the model with seasonal lags would have higher accuracy. A valid comparison can be made with respect to Kole and Dijk (2010)'s Markovian logit models, which also predicts between the two market states one week in advance. Kole and Dijk (2010)'s best model has an accuracy of 89.3%, which is slighly lower than our best model (91.7% accuracy) which use 1-month duration data to forecast the market regime one week in advance.

Despite the higher accuracy of our machine learning model, there are still several methods to improve the model accuracy in this research. First, when the minimum spanning trees were constructed, the minimization of their criterion did not converge, which suggested that the minimum spanning trees we have may not be the most efficient representations of the correlation networks. Second, some of the market regimes partitioned by our classification algorithms have extremely short duration such as one or two days, which actually reflect some of the Black Swain events such as the bust of the Subprime Mortgage bubble. However, because the ensuing market rallies and corrections are so close to those extreme market events, it is difficult to using approximately the same lagged data to predict multiple market regimes. Therefore, this paper focuses on only predicting the bull and bear markets. If these improvements can be properly implemented, then the model accuracy may increase accordingly.

Our work moves beyond previous work implementing both financial data and correlation networks, by being one of the first to introduce a use for these networks in market

classification. We have also created a framework by which a forecasting model for market regimes could be created. This has a number of applications, both theoretical and practical. Theoretically, it opens the door to research on the use of network theory in finance, and how more complex networks and analysis could reveal valuable information. Practically, our work could be used a number of ways. Similar work can be done to show how any subset relate to and influence a larger set, such as industries to the market, national GDP's to international growth, etc. Robust and early classification of market trends provide financial practitioners with early warning about hard times ahead, and the all-clear earlier than waiting for a trend to be established in the market itself. Therefore, with this application, financial practitioners will be able to act in advance of the market extremes and incorporate this forecast into their asset allocation strategies (Bernhart and Zagst, 2011). Expansion from classifying to forecasting provides similar use, and could also be used by governments and central banks to monitor the future health of their economies and financial markets. This in turn could potentially be linked to GDP/recession-forecasting efforts.

7 Acknowledgement

I would like to thank my thesis advisor, Dr. Michael Aguilar, for his mentorship and guidance. From providing the impetus for this research to aiding in the development of the econometric model, Dr. Aguilar has provided an immeasurable amount of aid and direction in this process. Without Dr. Aguilar's guidance, this thesis would not have been completed.

I would also like to thank my faculty advisor, Dr. Jane Fruehwirth for guiding me and the entire honor thesis class through the completion of this thesis. She provided the framework in which research is conducted, and with her guidance, I was able to remain on track throughout this entire process.

References

Ang, A., . B. G. (2002). International asset allocation with regime shifts. *The Review of Financial Studies*, 15(4):1137–1187.

Bernhart, G., H. S. N. M. N. M. and Zagst, R. (2011). Asset correlations in turbulent markets and the impact of different regimes on asset management. *Asia - Pacific Journal of Operational Research*, 28(1):1–23.

Bonanno, G., Caldarelli, G., Lillo, F., and Mantegna, R. N. (2003). Topology of correlation-based minimal spanning trees in real and model markets. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 68(4).

Degiannakisa, S., F. G. and Florosc, C. (2013). Oil and stock returns: Evidence from european industrial sector indices in a time-varying environment. *Journal of International Financial Markets, Institutions and Money*, 26:175–191.

Diebold, F. Lee., H. J. and Weinbach, G. (1993). Regime switching with time-varying transition probabilities. *Nonstationary Time Series Analysis and Cointegration*.

Fama, E. and French, K. (1992). The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465.

French, K. (2019). Kenneth r. french data library.

Guo, X., Zhang, H., and Tian, T. (2018). Development of stock correlation networks using mutual information and financial big data. *PLoS ONE*, 13(4):e0195941.

Hanna, A. (2018). A top-down approach to identifying bull and bear market states. *International Reviews of Financial Analysis*, 55:93–110.

Kazemilari, M. and Djauhari, M. A. (2015). Correlation network analysis for multi-dimensional data in stock market. *Physica A: Statistical Mechanics and Its Applications*, 429:62–75.

Kole, E. and Dijk, D. J. (2010). How to identify and predict bull and bear markets. *Paris December*.

Kruskal, J. (1964). Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27.

Kullmann, L., Kertész, J., and Kaski, K. (2002). Time-dependent cross-correlations between different stock returns: A directed network of influence. *Physical Review. E: Statistical, Nonlinear, and Soft Matter Physics*, 66(2):026125/6–026125.

Lunde, A. and Timmermann, A. (2004). Duration dependence in stock prices: An analysis of bull and bear markets. *Journal of Business Economic Statistics*, 22(3):253–273.

Mantegna, R. (1999). Hierarchical structure in financial markets. *The European Condensed Matter and Complex Systems*, 11(1):193–197.

Mantegna, R. N. and Stanley, H. E. (1999). *Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press.

Pierdzioch, C. and Risse, M. (2018). A machine-learning analysis of the rationality of aggregate stock market forecasts. *International Journal of Finance and Economics*, 23(4):642–654.

Sun, X., Wang, J., Yao, Y., Li, J., and Li, J. (2019). Spillovers among sovereign cds, stock and commodity markets: A correlation network perspective. *International Review of Financial Analysis.*

Upadhyay, A., B. G. . D. A. (2012). Forecasting stock performance in indian market using multinomial logistic regression. *Journal of Business Studies Quarterly*, 3(3):16–39.

Wee, M. and Yang, J. (2011). Order size, order imbalance and the volatility–volume relation in a bull versus a bear market. *Accounting Finance. Special Issue: Global Financial Crisis*, 52(1):145–163.

8 Appendix

Table 8: Partition Dates

Date	Regime	Date	Regime
3/3/78	Peak	4/14/00	Correction
9/11/78	Peak	9/1/00	Peak
11/14/78	Trough	4/4/01	Trough
10/5/79	Peak	5/21/01	Peak
11/7/79	Trough	9/21/01	Trough
2/13/80	Peak	1/4/02	Peak
3/27/80	Trough	7/23/02	Trough
11/28/80	Peak	8/22/02	Peak
9/25/81	Trough	10/9/02	Trough
11/30/81	Peak	11/27/02	Peak
3/8/82	Trough	3/11/03	Trough
5/7/82	Peak	10/9/07	Peak
8/12/82	Trough	3/10/08	Trough
10/10/83	Peak	5/19/08	Peak
7/24/84	Trough	10/10/08	Trough
8/25/87	Peak	10/13/08	Rally
10/19/87	Trough	10/27/08	Correction
10/21/87	Rally	11/4/08	Rally
10/26/87	Correction	11/20/08	Correction
11/2/87	Rally	1/6/09	Peak
12/4/87	Trough	3/9/09	Trough
10/9/89	Peak	4/23/10	Peak
1/30/90	Trough	7/2/10	Trough
7/16/90	Peak	4/29/11	Peak
10/11/90	Trough	10/3/11	Trough
10/7/97	Peak	5/21/15	Peak
10/27/97	Correction	8/25/15	Trough
7/17/98	Peak	11/3/15	Peak
8/31/98	Trough	2/11/16	Trough
9/23/98	Rally	1/26/18	Peak
10/8/98	Correction	2/8/18	Correction
7/16/99	Peak	9/20/18	Peak
10/15/99	Trough	12/24/18	Trough
3/24/00	Peak	6/20/19	Peak